Письма в ЭЧАЯ. 2013. Т. 10, № 5(182). С. 756–763

БАЗА ДАННЫХ АМИНОКИСЛОТНО-НУКЛЕОТИДНЫХ КОНТАКТОВ В КОМПЛЕКСАХ ДНК С БЕЛКАМИ СЕМЕЙСТВА ГОМЕОДОМЕНОВ

Т. И. Грохлина [а], *П. В. Зрелов* [б], *В. В. Иванов* [б],

Р. В. Полозов [в], *Ю. Н. Чиргадзе* [г], *В. С. Сивожелезов* [д]

[а] Институт математических проблем биологии РАН, Пущино, Россия

[б] Объединенный институт ядерных исследований, Дубна

[в] Институт теоретической и экспериментальной биофизики РАН, Пущино, Россия

[г] Институт белка РАН, Пущино, Россия

[д] Институт биофизики клетки РАН, Пущино, Россия

Анализ контактов аминокислот с нуклеотидами в интерфейсах комплексов белок–ДНК с целью поиска закономерностей ДНК-белкового узнавания — сложная задача, требующая одновременного анализа физико-химических характеристик этих контактов, позиций участвующих в контактах аминокислот и нуклеотидов в последовательностях белка и ДНК и консервативности этих контактов. Таким образом, необходимо систематизировать эти разнородные данные, для чего была разработана база данных аминокислотно-нуклеотидных контактов ANTPC (Aminoacid Nucleotide Type Position Conservation) на примере белков из семейства гомеодоменов. Показано, что она может быть использована для сравнения и классификации ДНК-белковых интерфейсов.

The analysis of amino acid–nucleotide contacts in interfaces of the protein–DNA complexes, intended to find consistencies in the protein–DNA recognition, is a complex problem that requires analysis of the physicochemical characteristics of these contacts, of the positions of the participating amino acids and nucleotides in the chains of the protein and the DNA, respectively, as well as conservatism of these contacts. Thus, those heterogeneous data should be systematized. For this purpose we have developed a database of amino acid–nucleotide contacts ANTPC (Amino acid Nucleotide Type Position Conservation) following the archetypal example of the proteins in the homeodomain family. We show that it can be used for comparison and classification of interfaces of the protein–DNA complexes.

PACS: 82.37.Rs

ВВЕДЕНИЕ

В течение достаточно длительного времени в литературе обсуждаются вопросы, связанные с проблемой узнавания ДНК определенными семействами белков. В частности, предметом бурной дискуссии было существование кода ДНК-белкового узнавания, выраженного через непосредственные контакты между аминокислотами и нуклеотидами [1–3]. Для различных видов комплексов ДНК с белками были получены некоторые закономерности, например, оказались характерными контакты Arg:Gua или Asn:Ade, однако в общем случае эти и другие закономерности оказались справедливыми только вероят-

ностно [4]. Попытки вывести точные правила узнавания не удались в том числе потому, что в предыдущих исследованиях рассматривались разнородные семейства комплексов ДНК с белком. По оценкам, сделанным с помощью базы данных SCOP [5], существует около двух тысяч ДНК-белковых комплексов известной трехмерной структуры, принадлежащих 207 семействам.

Для анализа мы выбрали семейство гомеодоменов. Интерфейсы ДНК–белок характеризуются положением вектора $C\alpha{-}C\beta$ каждой из ДНК-связывающих аминокислот относительно векторов нормалей к плоскостям пар оснований — «стерическими соотношениями» по Пабо и Неклюдовой [6]. Оказалось, что существует множество ориентаций аминокислот относительно узнаваемых пар оснований, что сильно затрудняет поиск общих правил ДНК-белкового узнавания. Эти ориентации как раз и определяют «стерические соотношения». Однако гомеодомены являются семейством белков, для которого вышеупомянутые «стерические соотношения» в интерфейсах сохранились в процессе эволюции в течение нескольких миллионов лет. Поэтому должен существовать определенный набор правил узнавания для всего семейства гомеодоменов. Кроме того, гомеодомены кодируются гомеобоксами — представителями одного из наиболее консервативных семейств генов [7]. На ранних стадиях развития гомеодомены контролируют морфогенез и органогенез эмбриона [8,9]. Интерфейсы гомеодомен–ДНК консервативны на протяжении 500 млн лет [10] и наблюдаются во всех эукариотах, имеющих предполагаемого общего предка [11–13]. Взаимодействия гомеодомен–ДНК в комплексах подробно рассмотрены в [14]. Все это побудило нас выбрать комплексы ДНК с белками именно этого семейства для выявления правил узнавания.

Ранее нами были детально изучены все контакты пяти комплексов гомеодомен–ДНК, полученных с помощью рентгеноструктурного анализа с высоким разрешением, и найдены как инвариантные, так и вариабельные контакты, а затем проанализированы контакты репрезентативного набора из 22 комплексов гомеодомен–ДНК. Основным объектом этого исследования были инвариантные контакты. Мы нашли позиционно-специфичный набор инвариантных контактов, который присутствует во всех структурах комплексов гомеодомен–ДНК, но отсутствует в комплексах ДНК с другими белками. Замечательно, что этот набор контактов является эволюционно консервативным для различных таксономических групп семейства гомеодоменов. Он включает один высококонсервативный контакт аспарагина с аденином и несколько позиционно-специфичных контактов фосфата с заряженными аминокислотными остатками. Мы предположили, что этот пространственный инвариант может считаться специфичным правилом узнавания при образовании комплексов гомеодоменов с операторной ДНК.

С целью проверки адекватности вышеуказанных закономерностей и обнаружения возможных новых решили создать базу данных ANTPC (Aminoacid Nucleotide Type Position Conservation) ДНК-белковых контактов по всем известным структурным данным семейства комплексов гомеодомен–ДНК.

1. ОПИСАНИЕ БАЗЫ ДАННЫХ ANTPC

На основе данных ЯМР и рентгеноструктурного анализа создана база данных ANTPC, содержащая сведения о контактах 68 комплексов факторов транскрипции семейства гомеодоменов с ДНК. В ней отражены сведения о типах взаимодействия и позициях контактов гомеодомен–ДНК в их первичных структурах.

В качестве полей базы данных определены идентификатор комплекса в Protein Data Bank (PDB) [15], с которым связана общая информация из этого банка данных — код цепи белка, образующего данный интерфейс, биологический вид, к которому он принадлежит, эмпирическое имя белка, название гена в геноме человека.

Поля таблицы контактов содержат следующую информацию:

— идентификатор комплекса в Protein Data Bank;

— номер нуклеотида в нумерации, где за первый нуклеотид принимается начало наиболее часто встречающегося узнаваемого гомеодоменами мотива TAAT [16];

— название нуклеотида;

— позицию контакта, т. е. информацию о номере аминокислоты, где за 0 принят номер первой аминокислоты, контактирующей с большим желобом узнаваемой ДНК [16];

— тип контакта (табл. 1);

— аминокислоту, с которой взаимодействует нуклеотид;

— степень консервативности контакта; в базе различаются четыре степени консервативности: «c» — консервативные, «m» — умеренно консервативные, «v» — вариабельные, «a» — отсутствие контакта.

Таблица 1. **Типы ДНК-белковых контактов, используемые в базе данных ANTPC**

b	Контакт нуклеотид–аминокислота через основание
b!	С образованием бидентатной водородной связи с основанием ДНК
p	Контакт нуклеотид–аминокислота через фосфат
s	Нуклеотид связывается с аминокислотой через сахар
:	Аминокислота может связываться с несколькими нуклеотидными основаниями (бифуркационная связь)

Таблица 2. **Бидентатные контакты с участием аминокислоты в 9-й аминокислотной позиции в семействе гомеодоменов**

idPDB	Цепь ДНК	Номер	Нуклеотид	Позиция контакта	Тип контакта	Аминокислота	Консервативность
1akh_A	1	2	G	9	b!	R	m
1b8i_B	1	2	G	9	b!	R	m
1lfu_P	1	2	G	9	b!	R	m
1puf_B	1	2	G	9	b!	R	m
1yrn_A	1	2	G	9	b!	R	m
2d5v_A	1	2	G	9	b!	R	m
2d5v_B	1	2	G	9	b!	R	m
2r5y_B	1	2	G	9	b!	R	m
2r5z_B	1	2	G	9	b!	R	m

Для удобства анализа созданы таблицы двух типов, где интерфейс для каждого из комплексов упорядочен либо по аминокислотным последовательностям узнающей спирали (табл. 2), либо по нуклеотидным последовательностям узнаваемой ДНК (табл. 3).

Поскольку узнающая спираль ориентирована относительно большого желоба единообразно во всем семействе гомеодоменов, то позиции аминокислот в последовательности задают их пространственное положение в интерфейсах.

Таблица 3. **Общая характеристика белок–ДНК комплексов, включенных в базу данных**

Код PDB	Цепь	Биологический вид	Эмпирическое имя	Имя гена в геноме человека
1ahd	P	*Drosophila melanogaster*	Antennapedia	HOX(A,B)(5,6)
1akhA	A	*Saccharomyces cerevisiae*	Mating type protein A1	Unknown
1akhB	B	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1apl	C	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1apl	D	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1au7	A	*Rattus norvegicus*	Pit1 POU homeodomain	POU1F1
1au7	B	*Rattus norvegicus*	Pit1 POU homeodomain	POU1F1
1b72	A	*Homo sapiens*	hoxb1	HOXB1
1b72	B	*Homo sapiens*	pbx1 preBcell leukemia homeobox	PBX1
1b8i	A	*Drosophila melanogaster*	Ultrabithorax	HOX(A,B)7
1b8i	B	*Drosophila melanogaster*	Extradenticle	PBX(14)
1cqt	A	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1cqt	B	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1du0	A	*Drosophila melanogaster*	Engrailed	EN2
1du0	B	*Drosophila melanogaster*	Engrailed	EN2
1e3o	C	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1fjl	A	*Drosophila melanogaster*	Paired	PAX7
1fjl	B	*Drosophila melanogaster*	Paired	PAX7
1gt0	C	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1hdd	C	*Drosophila melanogaster*	Engrailed	EN2
1hdd	D	*Drosophila melanogaster*	Engrailed	EN2
1hf0	A	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1hf0	B	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1ic8	A	*Homo sapiens*	HNF1A Hepatocyte nuclear factor 1a	HNF1A
1ic8	B	*Homo sapiens*	HNF1A Hepatocyte nuclear factor 1a	HNF1A
1ig7	A	*Mus musculus*	Msx1 homeodomain	MSX1
1jgg	A	*Drosophila melanogaster*	Evenskipped	EVX(1,2)
1jgg	B	*Drosophila melanogaster*	Evenskipped	EVX(1,2)*
1k61	A	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1k61	B	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1k61	D	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1le8	A	*Saccharomyces cerevisiae*	Mating type protein A1	Unknown
1le8	B	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1lfu	P	*Mus musculus*	pbx1 preBcell leukemia homeobox	PBX1
1mnm	C	*Saccharomyces cerevisiae*	mat alpha2	Unknown

Окончание табл. 3

Код PDB	Цепь	Биологический вид	Эмпирическое имя	Имя гена в геноме человека
1mnm	D	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1nk2	P	*Drosophila melanogaster*	VND/NK2 protein	NKX22
1o4x	A	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1oct	C	*Homo sapiens*	Oct1 POU Homeodomain	POU2F1
1puf	A	*Mus musculus*	hoxa9	HOXA9
1puf	B	*Homo sapiens*	pbx1 preBcell leukemia homeobox	PBX1
1xpx	A	*Drosophila melanogaster*	Homeoprospero domain	PROX1
1yrn	A	*Saccharomyces cerevisiae*	Mating type protein A1	Unknown
1yrn	B	*Saccharomyces cerevisiae*	mat alpha2	Unknown
1yz8	P	*Homo sapiens*	Pituitary homeobox 2	PITX(2,3)
1zq3	P	*Drosophila melanogaster*	bicoid protein	HOX(A,B)(5,6)
2d5v	A	*Homo sapiens*	one cut homeobox 1	ONECUT1
2d5v	B	*Homo sapiens*	one cut homeobox 1	ONECUT1
2h1k	A	*Homo sapiens*	pdx1 pancreatic and duodenal homeobox 1	PDX1
2h1k	B	*Homo sapiens*	pdx1 pancreatic and duodenal homeobox 1	PDX1
2h8r	A	*Homo sapiens*	HNF1 homeobox B	HNF1B
2h8r	B	*Homo sapiens*	HNF1 homeobox B	HNF1B
2hdd	A	*Drosophila melanogaster*	Engrailed	EN2
2hdd	B	*Drosophila melanogaster*	Engrailed	EN2
2hos	A	*Drosophila melanogaster*	Engrailed	EN2
2hos	B	*Drosophila melanogaster*	Engrailed	EN2
2hot	A	*Drosophila melanogaster*	Engrailed	EN2
2hot	B	*Drosophila melanogaster*	Engrailed	EN2
2r5y	A	*Drosophila melanogaster*	SCR Sex combs reduced	HOXA(47)B(47)C(46)D4
2r5y	B	*Drosophila melanogaster*	Extradenticle	PBX(14)
2r5z	A	*Drosophila melanogaster*	SCR Sex combs reduced	HOXA(47)B(47)C(46)D4
2r5z	B	*Drosophila melanogaster*	Extradenticle	PBX(14)
3cmy	A	*Homo sapiens*	PAX3 paired box 3	PAX3
3hdd	A	*Drosophila melanogaster*	Engrailed	EN2
3hdd	B	*Drosophila melanogaster*	Engrailed	EN2
9ant	A	*Drosophila melanogaster*	Antennapedia	HOX(A,B)(5,6)
9ant	B	*Drosophila melanogaster*	Antennapedia	HOX(A,B)(5,6)

Примечание. Классификация гомеодоменов по свойствам интерфейсов (инвариантность контактов).

2. ФУНКЦИИ БАЗЫ ДАННЫХ ANTPC

В данном разделе функции базы данных представлены примерами, получаемыми с ее помощью. Главной функцией ANTPC является сортировка контактов по их типу и позиции в последовательностях гомеодоменов и ДНК. Среди бидентатных водородных связей (b!) для кодирующей цепи ДНК наиболее часто встречаются контакты аденина в позиции 3 с аспарагином в позиции 5. Этот контакт проиллюстрирован рис. 1.

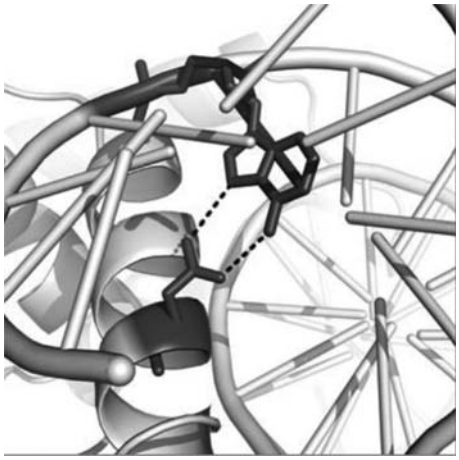

Рис. 1. Контакт аспарагина в 5-й позиции узнающей спирали с аденином в 3-й позиции узнаваемого мотива ДНК на примере комплекса с PDB кодом 1akh. В указанном комплексе эти позиции имеют номера 120 и 26 соответственно. Нумерация аминокислот в узнающей спирали определяется первым полярным контактом с ДНК, которому приписывается номер 0 [16], но не началом самой спирали

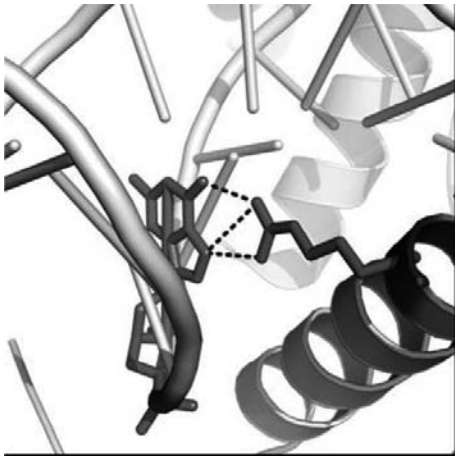

Рис. 2. Контакт аргинина в 9-й позиции узнающей спирали с аденином во 2-й позиции узнаваемого мотива ДНК на примере комплекса с PDB кодом 2r5z. В указанном комплексе эти позиции имеют номера 255 и 8 соответственно

Это один из самых консервативных контактов в комплексах гомеодоменов с ДНК. С помощью базы данных легко убедиться, что лишь два из включенных в нее комплексов не обладают бидентатным контактом в 5-й аминокислотной позиции — 1k61_D, у которого узнавание происходит со сдвигом на четыре аминокислоты, и 1le8_B, где в гомеодомене имеется искусственная мутация аспарагина-5 на аланин.

Бидентатные контакты встречаются также во второй нуклеотидной позиции, причем это всегда гуанин, взаимодействующий с аргинином в 9-й позиции (рис. 2).

Соответствующая выборка белок–ДНК комплексов из базы данных ANTPC представлена в табл. 2. Комплексы, имеющие такие контакты, образуют подсемейство гомеодоменов, кодируемых генами типа PBX, его гомологами и геном ONECUT1, а также гомеодоменом, специфичным для дрожжей (табл. 3). Однако гены PBX кодируют также и гомеодомены, не содержащие указанных контактов (табл. 3).

Аспарагин-адениновый контакт в 1-й (вместо 5-й, как в других случаях) аминокислотной позиции встречается лишь в вышеупомянутом комплексе 1k61_D (не показано). 1-я позиция узнающей спирали образует либо неполярный контакт с тимином в 1-й позиции узнаваемого нуклеотидного мотива, либо, в случае нахождения в ней полярной аминокислоты, с фосфатом в той же позиции (рис. 3).

Среди свойств интерфейсов важнейшим является консервативность входящих в него контактов. Мы обнаружили, что консервативным является, например, контакт триптофана во 2-й позиции аминокислотной последовательности с фосфатом нуклеотида также

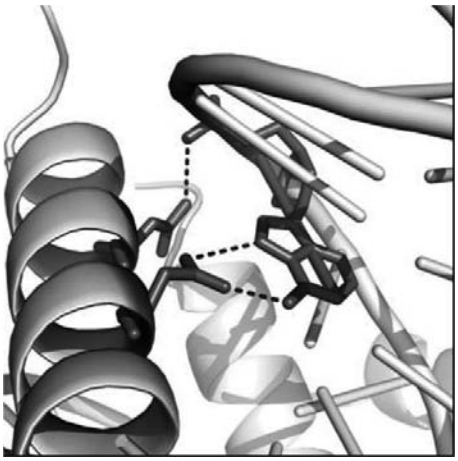

Рис. 3. Контакты двух аспарагинов в 5-м и 1-м положении, различающихся на четыре позиции (один виток спирали) в аминокислотной последовательности узнающей спирали. Они взаимодействуют с аденином в положении 3 нуклеотидного мотива и принадлежащим ему фосфатом, соответственно, на примере комплекса с PDB-кодом 1b8i, остатками аспарагина 254 и 250 и нуклеотидом аденин-21

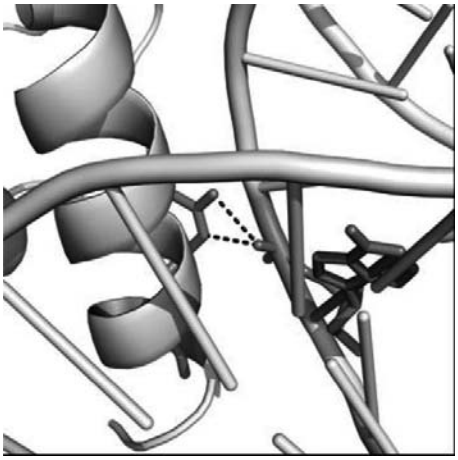

Рис. 4. Инвариантный контакт аргинина в 7-й позиции аминокислотной последовательности узнающей спирали с фосфатом нуклеотида в 5-й позиции узнаваемого мотива обратной цепи ДНК, на примере аргинина-122 и аденина-16 (комплементарного тимину-29) комплекса с PDB-кодом 1akh

во 2-й позиции, но нуклеотидной последовательности узнаваемого мотива ДНК. В одном случае этот триптофан заменен на фенилаланин. Тогда, как мы заметили ранее [17], в узнавании участвует ацетат-ион, контактирующий своей метильной группой с фенилаланином, а карбоксильной группой образующий водородную связь с фосфатом.

Еще более консервативным является контакт аргинина в 7-й позиции аминокислотной последовательности с фосфатом нуклеотида в 5-й позиции узнаваемого мотива обратной цепи ДНК (рис. 4). Это единственный консервативный контакт гомеодомена с обратной цепью ДНК.

База данных ANTPC позволяет устанавливать видовую специфичность или универсальность контактов. Мы обнаружили, что контакт серина в нулевой аминокислотной позиции с 7-м нуклеотидом обратной цепи присутствует у трех биологических видов: дрозофилы, мыши и человека, и не присутствует в гомеодоменах дрожжей. Это наблюдение и сравнительный анализ аминокислотных последовательностей гомеодоменов могут быть полезны при решении некоторых задач эволюции гомеодоменов.

ЗАКЛЮЧЕНИЕ

Наш подход состоит в том, чтобы разработать базы данных для отдельных семейств ДНК-узнающих белков, а затем объединить их по определенным правилам в единый банк данных контактов белок–ДНК. База данных ANTPC для семейства гомеодоменов позво-

ляет систематизировать разнородные данные, такие как позиции и физико-химические свойства всех контактов в интерфейсах белков данного семейства с ДНК. Кроме того, наша база данных позволяет решать задачи сравнения и классификации интерфейсов белок–ДНК. Все это трудно было бы сделать с использованием общепринятых в настоящее время методов и подходов к созданию аналогичных баз данных, включающих большое разнообразие семейств ДНК-узнающих белков. Решить указанные задачи удается именно потому, что было рассмотрено лишь одно семейство ДНК-узнающих белков.

Работа поддержана РФФИ, грант № 11-07-00374.

СПИСОК ЛИТЕРАТУРЫ

1. *Matthews B. W.* // Nature. 1988. V. 335. P. 294–295.
2. *Suzuki M. et al.* // Protein Eng. 1995. V. 8. P. 319–328.
3. *Choo Y., Klug A.* // Curr. Opin. Struct. Biol. 1997. V. 7. P. 117–125.
4. *Benos P. V., Lapedes A. S., Stormo G. D.* // Bioessays. 2002. V. 24. P. 466–475.
5. *Murzin A. G. et al.* // J. Mol. Biol. 1995. V. 247. P. 536–540.
6. *Pabo C. O., Nekludova L.* // J. Mol. Biol. 2000. V. 301. P. 597–624.
7. *Kalthoff K.* Analysis of Biological Development. N. Y.: McGraw-Hill, 1996.
8. *Svingen T., Koopman P.* // Sex Dev. 2007. V. 1. P. 12–23.
9. *Wigle J. T., Eisenstat D. D.* // Clin. Genet. 2008. V. 73. P. 212–226.
10. *Gehring W. J., Affolter M., Burglin T.* // Ann. Rev. Biochem. 1994. V. 63. P. 487–526.
11. *Derelle R. et al.* // Evol. Dev. 2007. V. 9. P. 212–219.
12. *Kissinger C. R. et al.* // Cell. 1990. V. 63. P. 579–590.
13. *Rubin G. M. et al.* // Science. 2000. V. 287. P. 2204–2215.
14. *Ledneva K. et al.* // Mol. Biol. 2001. V. 35. P. 647–659.
15. *Berman H. M. et al.* // Nucleic Acids Res. 2000. V. 28. P. 235–242.
16. *Chirgadze Yu. N. et al.* // J. Biomol. Struct. Dyn. 2009. V. 26. P. 687–700.
17. *Chirgadze Yu. N. et al.* // J. Biomol. Struct. Dyn. 2012. V. 29. P. 715–731.

Получено 30 ноября 2012 г.